Journal of Biotechnology 235 (2016) 132–138

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

Review

Synergism of proteomics and mRNA sequencing for enzyme discovery

Lukas Sturmberger[a,1], Paal W. Wallace[a,b,c,1], Anton Glieder[a,d], Ruth Birner-Gruenberger[a,b,c,*]

[a] Austrian Center of Industrial Biotechnology (ACIB), Petersgasse 14, 8010 Graz, Austria
[b] Medical University of Graz, Institute of Pathology, Research Unit Functional Proteomics and Metabolic Pathways, Stiftingtalstrasse 24, 8010 Graz, Austria
[c] Omics Center Graz, BioTechMed-Graz, Stiftingtalstrasse 24, 8010 Graz, Austria
[d] TU Graz, Institute of Molecular Biotechnology, NAWI Graz, Petersgasse 14, 8010 Graz, Austria

ARTICLE INFO

Article history:
Received 9 November 2015
Received in revised form 7 December 2015
Accepted 14 December 2015
Available online 18 December 2015

Keywords:
Enzyme discovery
Proteomics
Transcriptomics
Database
Biotechnology

ABSTRACT

Enzyme catalyzed processes are increasingly complementing chemical manufacturing as new enzymes are being discovered. Although, many industrially applied biocatalysts have been identified by functional screenings technological advances in the omics fields have created a different path to access novelty. Here we describe how omics technologies, especially proteomics and transcriptomics, can complement each other in the aim of finding new enzymatic functions. Special emphasis is laid on how mRNA sequencing Zcan improve proteomic experiments by allowing the generation of high quality protein sequence databases, which subsequently facilitates protein identification.

Contents

Abbreviations: mRNA, messenger ribonucleic acids; cDNA, complementary desoxyribonucleic acids; MALDI-TOF, matrix-assisted laser desorption/ionization followed by time of flight; LC–MS/MS, liquid chromatography tandem mass spectrometry; NCBI, National Center for Biotechnology Information; EMBL, European Molecular Biology Laboratory; DDBJ, DNA Data Bank of Japan; IINSDC, International Nucleotide Sequence Database Collaboration; JGI, Joint Genome Institute; rRNA, ribosomal ribonucleic acids; ESI, electrospray ionization; EST, expressed sequence tag; FDA, food and drug administration; SL, sophorolipid; SILAC, stable isotope labeling by amino acids in cell culture; TCA, tricarboxylic acid.

* Corresponding author.
E-mail address: ruth.birner-gruenberger@medunigraz.at (R. Birner-Gruenberger).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.jbiotec.2015.12.015

1. Introduction

Enzymes have over the course of the last decade established themselves as useful alternatives to classical organo-and metallo-catalysis (Anastas and Warner, 1998; Bornscheuer et al., 2012; Illanes et al., 2012; Meyer, 2011). They are increasingly employed for the manufacturing of a diverse array of chemical products such as pharmaceuticals, agrochemicals, bulk chemicals or biofuels on preparative as well as industrial scale (Erickson et al., 2012). This focus also applies to biological degradation of products, such as xenobiotics, plastics, biomass and unwanted components in wastewater (Cammarota and Freire, 2006; Eberl et al., 2008; Janssen et al., 2005; Kullman and Matsumura, 1996; Müller et al., 2005; Rabinovich et al., 2004; Ribitsch et al., 2012). So far about 200 enzymes are industrially used (Li et al., 2012), with the majority of those being hydrolytic enzymes and oxidoreductases (Faber, 2011). The importance of enzyme-catalyzed synthesis can also be visualized in economic terms. The global enzyme market in 2010 was worth 3.3 billion US dollars and is expected to reach 4.4 billion dollars by the end of 2015 (Li et al., 2012; Sarrouh et al., 2012).

Due to the need for enzyme-catalyzed processes, the biotechnological sector is on a constant quest for new sources of enzymes. The majority of discovered and applied enzymes derive from a limited number of cultivable laboratory organisms, mainly of fungal or bacterial origin (Robertson and Steer, 2004). Estimates show that less than 1% of all microbes are readily cultivable due to unknown cultivation conditions, slow growth behavior or the absence of essential nutrients supplied by other organisms (Amann et al., 1995; Ekkers et al., 2012). This fact presents both a challenge and an opportunity for the discovery of novel enzymes and functions. Most industrially applied enzymes have been found by functional screenings of metagenomic and genomic libraries (Ferrer et al., 2009; Uchiyama and Miyazaki, 2009). The advantages of functional screens lie in the immediate identification of enzyme activity whereas methods based solely on sequence only give predictions about enzyme function derived from annotations. On the other hand, the complex makeup of eukaryotic genomes containing many layers of regulatory elements drastically increases library size, making functional expression and screening of these libraries very challenging. Even though functional screening of bacterial and to a lesser extent eukaryotic genomes and metagenomes have been most successful, mere *in silico* methods such as structure guided (Steinkellner et al., 2014) and sequence similarity-based approaches have also received attention in recent years. The latter approaches are limited to the discovery of new genes with high similarities to already deposited sequences, making it impossible to discover truly novel enzymatic functions. They nevertheless represent a powerful tool by eliminating cost-and labor-intensive wet-laboratory time (Behrens et al., 2011).

Although genomic screenings have been successful in identifying new enzymes, recent technological advances in the field of transcriptomics and proteomics have made these attractive tools for further discoveries. A major disadvantage of metagenomic or genomic libraries is the high percentage of genomic DNA comprising of non-coding regions (Deutsch and Long, 1999), which need to be removed in order to produce functional proteins and which unnecessarily increase library size. In addition to problems related to incorrect positioning of the gene relative to its promoters, non-spliceable introns and different codon usage by different organisms, posttranslational modifications are challenging when expressing these genes in hosts other than the original organism. (Behrens et al., 2011). By applying a transcriptomic or proteomic approach, some of these disadvantages are circumvented. Using transcriptomics instead of genomics the majority of non-coding DNA elements can be removed thereby reducing library size and avoiding non-functional genes arising due to incomplete gene-splicing recognition. Furthermore, the positioning of cDNA transcript sequences in relation to core promoter regions allows for a higher likelihood of successful transcription and translation compared to the majority of sheared genomic DNA fragments. By using transcriptomics rather than genomics, it is also possible to investigate dynamic spatio-temporal gene expression patterns. Therefore, transcriptomics can capture differential gene expression arising due to changes in environmental factors such as the presence of certain chemicals or shifts in cultivation conditions. This dynamic can also be monitored using proteomics. Quantitative proteomics has an advantage over transcriptomics as mRNA abundance does not accurately reflect protein abundance (Gygi et al., 1999; Zhang et al., 2014) and only proteomics can capture posttranslational modifications. Proteomic analysis moreover has the benefit of analysis directly performed on proteins expressed by the original organism. This circumvents problems related to transcription and translation of the gene and the potential subsequent posttranslational modification of the protein in different host organisms. Proteomics additionally allows the direct elucidation of subcellular localization whereas transcriptomics and genomics rely on predictions. Proteins can be localized to different subcompartments, such as the cytosol, peroxisomes, mitochondria, endoplasmatic reticulum, Golgi vesicles, lipid droplets, lysosomes, and membranes or to the extracellular space. This information provides valuable cues about the conditions under which the proteins are functional and stable. In turn, this allows for optimization of the conditions for functional screenings, which can increase the number of hits. One of the most direct ways to discover enzymes is through activity based proteomics which relies on enzyme class specific probes for simultaneous identification of individual enzymatic activities sharing the same reaction mechanism (Cravatt et al., 2008; Schittmayer and Birner-Gruenberger, 2012). This does however require the development of appropriate probes.

2. Enzyme discovery workflow

The most commonly used proteomic approach for enzyme discovery is bottom-up shotgun proteomics. In this approach proteins are enzymatically digested into peptides which are analyzed either by tandem mass spectrometry after a separation on a liquid chromatography system and electrospray ionization (ESI–LC–MS/MS) or by matrix-assisted laser desorption/ionization followed by time of flight mass spectrometry (MALDI–TOF–MS) (Fig. 1). Identification of proteins is based on the comparison of the measured mass to charge ratios (m/z) of the peptides and their fragments to the respective m/z values for all theoretical peptides and fragments thereof present in the database, which are generated by in silico digestion of the database with the same enzyme. The comparisons are automatically performed by search programs such as MASCOT (Pappin et al., 1999), SEQUEST (Eng et al., 1994), MaxQuant (Cox and Mann, 2008), MSAmanda (Dorfer et al., 2014), OMSSA (Geer et al., 2004), X!Tandem (Craig and Beavis, 2004), COMET (Eng et al., 2013), MS–GF+ (Kim and Pevzner, 2014) and MyriMatch (Tabb et al., 2008). All these search engines infer the presence of a protein from identification of at least one peptide that is unique to this protein (within the used database). This however implies that for a protein to be correctly identified it needs to be present in the database. Although the possibility of performing *de-novo* sequencing of proteins exists (Hughes et al., 2010), and could circumvent the problem of a protein not being present in the database, it is time and labor intensive and requires large amounts of highly purified proteins.

The most direct way to construct a high quality database for peptide mapping is by building it from the same sample used for protein identification by either genome or transcriptome sequencing (Fig. 1). For cultivable and, especially, model organisms, genome

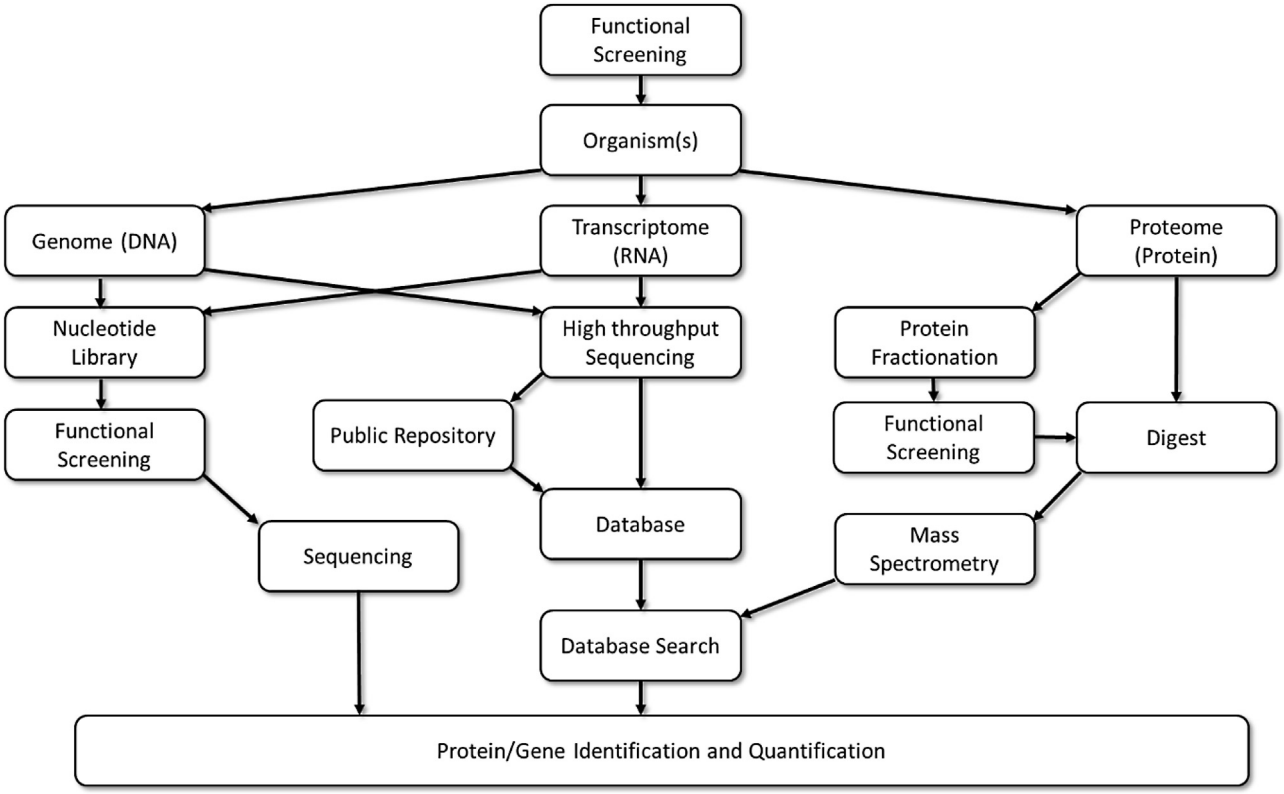

Fig. 1. Workflow of omics experiments for the discovery of novel enzymes. Functional screenings lead to the identification of organisms, tissues or other samples of interest. These can be analysed omics technologies to identify the gene/protein of interest. Genome and transcriptome sequence data can be integrated into the proteomics workflow, by generating databases, which can improve the number of correct protein identifications from mass spectrometric data. This facilitates the access to novel enzymes by recombinant expression in microbial hosts.

sequences (and even curated protein databases like SwissProt) are publicly available and present an easy way to obtain a high quality database. Among the most prominent suppliers of sequence databases are the National Center for Biotechnology Information (NCBI), the European Molecular Biology Laboratory (EMBL) and the DNA Data Bank of Japan (DDBJ), which are all integrated through the International Nucleotide Sequence Database Collaboration (INSDC). Other important public database resources are the Uniprot Knowledge database (UniProtKB) (Consortium, 2015) or the Joint Genome Institute's genome portal (JGI) (Nordberg et al., 2014). However, if public databases lack sequence information of the target organism, which is the case for most complex environmental samples, custom database creation is necessary. If nucleotide sequencing is unfeasible, a reasonable approach is to compile a database consisting of publically available sequences from (more or less) related organisms. This has been shown to be less reliable than using databases based on the target organism alone (Bräutigam et al., 2008) and only allows prediction of protein function but not of the exact sequence, preventing cloning and expression of the protein. Also, by querying against a database not containing the target organism's sequences, potential novel enzymes, that lack any homology with other proteins, cannot be identified.

In order to minimize the loss of information it is advisable to create a custom protein database translated from the sample's transcriptome. Advances within the field of RNA sequencing have allowed for the analysis of entire meta-transcriptomes and made them available for construction of protein databases. Particularly developments in cDNA generation protocols and increasing sequencing capacity (McGettigan, 2013) has made transcriptomics more accessible to a wider audience being offered as an affordable service by several companies. The preparation of cDNA involves three major steps: (1) the extraction of RNA from cultured organisms or environmental samples, (2) the removal of rRNA and normalization to enrich for rare transcripts and (3) the reverse transcription to generate cDNA, which subsequently is sequenced. Although considerable progress has been made in assembly of RNAseq data methods, analysis still poses some challenges. Among these is the misalignment of reads to closely related genes (such as sequences coding for isoenzymes), also termed transcript shadowing (McGettigan, 2013) and the fact that many *de novo* assembly algorithms are still highly memory intensive and therefore require access to advanced computing facilities (Grabherr et al., 2011).

Proteomic and metaproteomic screenings have been performed on both cultured microorganisms as well as environmental samples with the purpose of discovering novel enzymatic functions or for determining optimal culture conditions leading to e.g. optimal degradation of biomass or phosphorus removal. Here we describe a few selected examples of enzyme or whole pathway discovery where either a protein database from a public repository or a genome or transcriptome sequence has been obtained prior to proteomic analyses. Moreover, we discuss the advantages offered by combining proteomics with protein sequence databases derived from mRNA sequencing for enzyme discovery in more depth.

3. Pathway discovery for production of secondary metabolites

As the chemical synthesis of many useful compounds, like alkaloids or biosurfactants, can be long and complicated, using organisms that naturally synthesize these secondary metabolites can be very beneficial. However, to allow the heterologous production, the knowledge of the enzymatic steps involved is essential.

Here we present a few studies employing proteomic screenings using different types of protein sequence databases for this purpose and discuss the advantages and drawbacks of these databases.

3.1. Polyketide synthesis

One example, in which the availability of a high quality protein database could have assisted protein identification in a more direct way, is the elucidation of a non-ribosomal peptide synthetase (NRPS) and a polyketide synthase gene cluster from an unsequenced *Bacillus* sp. isolated from a soil sample. By exploiting the large size of these protein complexes, Evans et al. were able to identify peptides which served as the basis for PCR primer design (Evans et al., 2011). This approach allowed them to amplify the gene cluster containing several enzymes important for the synthesis of koranimine, a secondary metabolite, which falls into a family of compounds containing e.g. known mycotoxins and immunosuppressants. The gene cluster discovered was not highly related to any other NRPS cluster present in GenBank (the NCBI genetic sequence database) (Benson et al., 2013) at the time of analysis. Therefore, while publically available databases would most likely not have yielded protein identification an improved protein sequence database derived from mRNA sequencing could have allowed the direct identification of the protein sequence in the gel band without the need to perform difficult PCR amplifications using degenerate primers derived from amino acid sequences.

3.2. Alkaloid production

Papaver somniferum (opium poppy) is the only commercial source of the medically important alkaloids noscapine, morphine, codeine, papaverine and sanguinarine (Berenyi et al., 2009). The biosynthesis of these alkaloids have been studied extensively (Hagel and Facchini, 2013). Here we summarize the employed proteomic screening approaches and discuss how their outcome was improved by the use of protein sequence databases derived from Expressed Sequence Tag (EST, i.e. a short cDNA read) sequencing.

The first proteomic approach to detect and localize enzymes responsible for alkaloid biosynthesis was performed on the latex of the opium poppy. Latex proteins were extracted and separated by 2D-gel electrophoresis prior to in gel digestion, and peptide sequencing of picked gel spots (Decker et al., 2000). The resultant MS/MS data were queried against all the present sequences in the NCBI non-redundant protein database, as no *P. somniferum* specific database was available at the time. Of the 75 excised gel spots, 69 contained peptides that could be assigned to homologous proteins with known functions based on sequence-similarity to other organisms. Among the identified proteins was codeine reductase, which is known to be involved in morphine synthesis, confirming the presence of this enzyme in latex. Proteomics was also used in combination with more conventional approaches for enzyme discovery: Peptide sequencing and homology searches in combination with sequence-similarity based PCR-primers lead to the identification of two methyl-transferases involved in alkaloid biosynthesis (Ounaroon et al., 2003).

Similarly, Jacobs et al. (2005) were challenged with a lack of sequence resources for database construction in their quest for novel genes for the biosynthesis of alkaloids. They performed MALDI–MS/MS analyses of samples from the medicinal plant *Catharanthus roseus* and due to poor genome and proteome sequence information (only 27 deposited protein sequences in SwissProt at the time of analysis) they had to rely on the NCBInr database restricted to *Viridiplantae*. 58 different proteins, among these several enzymes from the alkaloid biosynthesis pathways, were identified. Of these, only one was a *C. roseus* protein. Thus, without access to specific sequence databases for *C. roseus* the huge potential of the peptide sequence data set could not be fully explored.

More recently, the availability of EST databases specific for opium poppies allowed for better identification of proteins involved in plant defense responses and synthesis of antimicrobial alkaloids in *P. somniferum* (Zulak et al., 2009). Latex proteins separated by 2D-gel electrophoresis were identified by LC–MS/MS and MALDI–TOF–MS and searching against both the NCBInr protein database restricted to *Viridiplantae* containing 381,872 sequences and a translated EST database containing 22,036 sequences. Of 219 identified proteins, 29% were found using the sequences contained only within the EST database while another 12% could be identified from sequences contained within both databases. In this study, transcriptomics was moreover included as a complementary method to quantify transcript levels under elicitor induced and non-induced conditions showing the induction of six known alkaloid biosynthetic enzymes.

As deeper sequencing of the transcriptomes was performed the size of the EST databases used for proteomic screenings increased, and 1004 peptides and polypeptides could be identified from opium poppy cell lines (Desgagné-Penix et al., 2010). Again, the search results of the public NCBInr database restricted to *Viridiplantae* were compared to those obtained by using an EST database (427,369 sequences). While searching the NCBInr database yielded only 288 identified peptides or polypeptides, querying against the opium poppy specific EST database resulted in the identification of 1004 peptides and polypeptides. This example clearly demonstrates the advantage of complementing proteomic studies with the use of organism specific EST databases.

3.3. Isoprenoid production

In a similar example, enzymes responsible for isoprenoid biosynthesis in tomato trichrome were discovered by a combination of mRNA sequencing and proteomics of the same organism. Schilmiller et al. (2010) sequenced the transcriptome (mRNA) and subsequently translated it into a protein sequence database which was then used for proteomic screening. This synergistic approach enabled the identification of 1552 proteins including all eight proteins involved in the production of the precursors needed for isoprenoid biosynthesis. Moreover, it led to the discovery of a previously uncharacterized tomato trichome sesquiterpene synthase involved in the synthesis of the anti-inflammatory compounds β-caryophyllene (approved as a food additive by the FDA) and α-humulene.

3.4. Sophorolipid production

Another prominent example for the synergism of different omics technologies is the discovery of enzymes involved in sophorolipid (SL) production. The yeast *Starmerella bombicola* (formerly *Candida bombicola*) was found to produce SL (Spencer et al., 1970) which have been suggested as environmentally friendly biosurfactants for the use in fields ranging from cleaning agents to stabilization of nanoparticles (Basak et al., 2013). Therefore, people have focused on the discovery of the genes necessary for the SL biosynthesis pathway. Initially, genes were identified one by one using sequence homology searches combined with PCR and gene walking, as no genome sequence was known (Saerens and Soetaert, 2011; Saerens et al., 2011; Van Bogaert et al., 2007). In 2013 the genome of *S. bombicola* was sequenced (Ciesielska et al., 2013) allowing for more comprehensive studies of the organism. This quickly led to the identification of a gene cluster containing all proteins known to be involved in SL synthesis (Van Bogaert et al., 2013). The cluster also contained other genes, one of which was found to encode a necessary SL transporter. The first large scale pro-

teomic study on *S. bombicola* identified and quantified 615 proteins. Quantitation was performed both by RNAseq and by a quantitative proteomic (stable isotope labeling by amino acids in cell culture (SILAC)) experiment, where the exponential and early stationary growth phases were compared. The authors reported a simultaneous production of all known proteins involved in SL biosynthesis (Ciesielska et al., 2013), which was followed up by an exoproteome analysis where the lactone esterase responsible for the final step of SL synthesis (lactonization) was identified (Ciesielska et al., 2014; Saerens et al., 2015). This example demonstrates the advantage of obtaining a genome sequence and combining RNAseq and quantitative proteomics for rapid identification of unknown enzymes, which is relatively easy and cheap for a cultivable organism.

4. Enzyme discovery

The importance of using a protein database containing the target protein or a very close homologue for the discovery of new enzymes, especially from uncultivable isolated organisms and environmental samples containing complex microbial communities, became similarly apparent in several studies published in recent years.

4.1. Isolated organisms

While searching for cellulases and xylanases Amore et al. isolated microorganisms from three different areas in the Western Ghat region of India (Amore et al., 2015). By analyzing eight different microorganisms obtained by cultivation, they were able to show xylanase activity in a *Bacillus amyloliquefaciens* XR44A strain. Subsequent HPLC–MS/MS analysis of peptides derived from a native gel band harboring xylanase activity and search of the LC–MS/MS data against the NCBInr database resulted in the identification of an endo-1,4-beta-xylanase. However, the protein identified in NCBInr was from *Paenibacillus macerans*, an organism not even belonging to the same family as the source organism (*Bacillaceae*). Thus, although the authors were able to identify an endo-1,4-beta-xylanase it is probably only a close homologue of the actual protein present in the sample, which could not be identified due to the absence of its protein sequence in the database. Since the analysis did not yield the actual sequence cloning and recombinant expression is restricted.

The same problem was faced by (Tiwari et al., 2014). After identifying hydrolytic activities in the secretome of the phytopathogenic fungus *Phoma exigua*, LC–MS/MS analysis was performed to identify potential protein targets with glycosyl hydrolase activity. Since the *P. exigua* genome was not available, searches were conducted against the NCBInr database restricted to fungal sequences and resulted in the identification of 33 different homologous proteins. The majority of these proteins were annotated as glycosyl hydrolases, but none of them was actually derived from *P. exigua* and therefore the exact sequence of the full-length protein was not accessible by the employed approach, once again restricting the further use of the obtained sequences.

Kirsch et al. (2012) studied plant cell wall degradation in the leaf beetle gut. The degradation is proposed to be mediated by enzymes secreted by the beetle itself. Thus, screenings of the gut proteome and transcriptome were performed to identify the plant cell wall degrading enzymes. Gel electrophoresis combined with activity assays resulted in gel bands with activity towards cellulose, pectin or xylan. To create an appropriate protein database and to identify potential enzymes, a metatranscriptomic analysis was performed, where several tissues were sampled during two developmental stages of the beetle and while being subjected to different environmental stress factors. The combined date resulted in an EST database with 644,940 entries, which was concatenated with the NCBInr database. The proteomic screening of the peptides from the active gel bands derived from the gut content identified 13 proteins from the beetle with a putative plant cell wall degrading enzymatic function, whereas the transcriptome of the same sample suggested 19 putative plant cell wall degrading enzymes. One may speculate that this difference might arise due to the higher specificity of the proteomic screen, low translation rates or protein stability of some candidates or wrong *in silico* functional assignment on the RNA level.

4.2. Microbial communities

With the aim of examining the suitability of activated zeolite as a carrier for microorganisms in anaerobic digestion processes, Weiß et al. performed LC–MS/MS of hydrolytically active protein bands after batch fermentation of grass silage (Weiß et al., 2013). Alongside this, they analyzed single strand conformation polymorphisms (SSCP) of the total bacterial community based on bacterial and archaeal 16S rRNA. By searching the LC–MS/MS data against the NCBInr database they were successful in identifying 36 biomass degradation associated enzymes, mainly from organisms in the *Paenibacillaceae* family. Meanwhile, predominantly species from the *Clostridium*, *Methanoculleus* and *Pseudomonas* families were shown to be the major organisms on activated zeolite by SSCP analysis. This can be interpreted in two ways: One possibility is that the specific proteomic analysis of enzymes highly complemented the 16S rRNA analysis of the total bacterial community. Since a small subset of the organisms present (*Paenibacillaceae* family) appeared to be responsible for the secretion of most of the hydrolytic enzymes, while not being among the most dominant organisms in terms of numbers. A second possible interpretation is that the protein identifications were assigned to the *Paenibacillaceae* family because the actual organisms that the proteins originate from were not present in the protein sequence database. In this case, the first interpretation seems to be more likely, since proteins from organisms more closely related to the organisms identified by SSCP than to the *Paenibacillaceae* family were present in the NCBInr database at the time.

The next example also shows that creation of custom protein databases by including relevant sequences into the analysis can improve the overall success in identification of novel enzymes. This approach of using focused databases was employed by Schneider et al. when querying their obtained peptide MS/MS spectra against a database compiled from both, a farm silage soil metagenome and the UniRef100 database (a clustered set of sequences from the UniProtKB) (Schneider et al., 2012). Using this metaproteomic approach they could show that on the one hand litter microbial communities differ between sampling sites and seasons and that on the other hand fungi are the main producers of litter-degrading enzymes. The majority of enzymes identified by this approach belonged to the group of cellulases, phosphatases, xylanases and lipases.

An even higher degree of specificity in database creation was implemented by a metaproteomic study on wastewater sludge. Wilmes et al. used a compilation of metagenomes derived from three different wastewater sludge sites to elucidate the enzymatic functions related to biological phosphorus removal. The use of these specific databases allowed them to identify enzymes involved in fatty acid oxidation and polyhydroxyalkanoate synthesis, glycogen degradation, TCA cycle, phosphate bioenergetics and stress response (Wilmes et al., 2008). The metagenomes, however, were generated from wastewater sludge from other sites than the metaproteomic study. Thus, although these databases proved very useful for enzyme identification, the list of protein identifications

may not be complete since the microbial composition of the sludge from the different sites may not be identical.

5. Conclusion

Proteomic approaches for gene/protein discovery depend on the usage of appropriate protein databases containing the sequences of the proteins present in the sample. Sequences of homologous proteins obtainable from public repositories have been used as databases instead, but the resulting data may be less comprehensive as demonstrated by the examples described above.

In the case of the discovery of enzymes involved in SL biosynthesis in *S. bombicola* (Ciesielska et al., 2013), the acquisition of a genomic sequence improved protein identifications. Similarly, the generation of EST databases by transcriptomic analysis accelerated enzyme discovery significantly as demonstrated by the examples of alkaloid biosynthesis in *P. somniferum* (Desgagné-Penix et al., 2010; Zulak et al., 2009) and plant cell wall degradation in leaf beetles (Kirsch et al., 2012). Kirsch et al.'s study on leaf beetles also demonstrates the benefits of transcriptomic sequence data both on its own and in its use to improve the protein database by combining the generated EST database with the NCBInr database. This approach of integrating mRNA sequencing data into protein databases would also be beneficial to environmental samples where the microbial composition (and thus potential proteome) is not known. This will allow the identification of a higher number of full-length protein sequences and their true origin rather than relying on close homologues found in other databases. With the reduced cost and increasing sequencing capacities proteomic approaches benefit from the generation of improved protein sequence databases derived from mRNA sequencing experiments. Furthermore one can expect that reanalysis of already acquired proteomic data with newly available improved protein sequence databases will yield valuable new information at very low cost.

Acknowledgements

This work was supported by the Austrian Science Fund (FWF) doctoral school "DK Metabolic and Cardiovascular Disease" (W1226), the Federal Ministry of Economy, Family and Youth (BMWFJ), the Federal Ministry of Traffic, Innovation and Technology (bmvit), the Styrian Business Promotion Agency SFG, the Standortagentur Tirol and ZIT-Technology Agency of the City of Vienna through the COMET-Funding Program managed by the Austrian Research Promotion Agency FFG.

References

Amann, R.I., Ludwig, W., Schleifer, K.H., 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Rev. 59, 143–169.

Amore, A., Parameswaran, B., Kumar, R., Birolo, L., Vinciguerra, R., Marcolongo, L., Ionata, E., La Cara, F., Pandey, A., Faraco, V., 2015. Application of a new xylanase activity from *Bacillus amyloliquefaciens* XR44A in brewer's spent grain saccharification. J. Chem. Technol. Biotechnol. 90, 573–581.

Anastas, P.T., Warner, J.C., 1998. Green Chemistry: Theory and Practice. University Press, Oxford.

Basak, G., Das, D., Das, N., 2013. Dual role of acidic diacetate sophorolipid as biostabilizer for ZnO nanoparticle synthesis and biofunctionalizing agent against *Salmonella enterica* and *Candida albicans*. J. Microbiol. Biotechnol. 24, 87–96.

Behrens, G.a., Hummel, A., Padhi, S.K., Schätzle, S., Bornscheuer, U.T., 2011. Discovery and protein engineering of biocatalysts for organic synthesis. Adv. Synth. Catal. 353, 2191–2215.

Benson, D.a., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2013. GenBank. Nucleic Acids Res. 41, D36–D42.

Berenyi, S., Csutoras, C., Sipos, A., 2009. Recent developments in the chemistry of thebaine and its transformation products as pharmacological targets. Curr. Med. Chem. 16, 3215–3242.

Bornscheuer, U.T., Huisman, G.W., Kazlauskas, R.J., Lutz, S., Moore, J.C., Robins, K., 2012. Engineering the third wave of biocatalysis. Nature 485, 185–194.

Bräutigam, A., Shrestha, R.P., Whitten, D., Wilkerson, C.G., Carr, K.M., Froehlich, J.E., Weber, A.P.M., 2008. Comparison of the use of a species–specific database generated by pyrosequencing with databases from related species for proteome analysis of pea chloroplast envelopes. J. Biotechnol. 136, 44–53.

Cammarota, M.C., Freire, D.M.G., 2006. A review on hydrolytic enzymes in the treatment of wastewater with high oil and grease content. Bioresour. Technol. 97, 2195–2210.

Ciesielska, K., Li, B., Groeneboer, S., Van Bogaert, I., Lin, Y.C., Soetaert, W., Van De Peer, Y., Devreese, B., 2013. SILAC-based proteome analysis of *Starmerella bombicola* sophorolipid production. J. Proteome Res. 12, 4376–4392.

Ciesielska, K., Van Bogaert, I.N., Chevineau, S., Li, B., Groeneboer, S., Soetaert, W., Van de Peer, Y., Devreese, B., 2014. Exoproteome analysis of *Starmerella bombicola* results in the discovery of an esterase required for lactonization of sophorolipids. J. Proteomics 98, 159–174.

Consortium, T.U., 2015. UniProt: a hub for protein information. Nucleic Acids Res. 43, D204–D212.

Cox, J., Mann, M., 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. 26, 1367–1372.

Craig, R., Beavis, R.C., 2004. TANDEM: matching proteins with tandem mass spectra. Bioinformatics 20, 1466–1467.

Cravatt, B.F., Wright, A.T., Kozarich, J.W., 2008. Activity-based protein profiling: from enzyme chemistry to proteomic chemistry. Annu. Rev. Biochem. 77, 383–414.

Decker, G., Wanner, G., Zenk, M.H., Lottspeich, F., 2000. Characterization of proteins in latex of the opium poppy (*Papaver somniferum*) using two-dimensional gel electrophoresis and microsequencing. Electrophoresis 21, 3500–3516.

Desgagné-Penix, I., Khan, M.F., Schriemer, D.C., Cram, D., Nowak, J., Facchini, P.J., 2010. Integration of deep transcriptome and proteome analyses reveals the components of alkaloid metabolism in opium poppy cell cultures. BMC Plant Biol. 10, 252.

Deutsch, M., Long, M., 1999. Intron–exon structures of eukaryotic model organisms. Nucleic Acids Res. 27, 3219–3228.

Dorfer, V., Pichler, P., Stranzl, T., Stadlmann, J., Taus, T., Winkler, S., Mechtler, K., 2014. MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. J. Proteome Res. 13, 3679–3684.

Eberl, A., Heumann, S., Kotek, R., Kaufmann, F., Mitsche, S., Cavaco-Paulo, A., Gübitz, G.M., 2008. Enzymatic hydrolysis of PTT polymers and oligomers. J. Biotechnol. 135, 45–51.

Ekkers, D.M., Cretoiu, M.S., Kielak, A.M., van Elsas, J.D., 2012. The great screen anomaly—a new frontier in product discovery through functional metagenomics. Appl. Microbiol. Biotechnol. 93, 1005–1020.

Eng, J.K., Jahan, T.A., Hoopmann, M.R., 2013. Comet: an open-source MS/MS sequence database search tool. Proteomics 13, 22–24.

Eng, J.K., McCormack, A.L., Yates, J.R., 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. 5, 976–989.

Erickson, B., Nelson, W., inters, P., 2012. Perspective on opportunities in industrial biotechnology in renewable chemicals. Biotechnol. J. 7, 176–185.

Evans, B.S., Ntai, I., Chen, Y., Robinson, S.J., Kelleher, N.L., 2011. Proteomics-based discovery of koranimine, a cyclic imine natural product. J. Am. Chem. Soc. 133, 7316–7319.

Faber, K., 2011. Biotransformations aid organic chemists. ChemInform.

Ferrer, M., Beloqui, A., Timmis, K.N., Golyshin, P.N., 2009. Metagenomics for mining new genetic resources of microbial communities. J. Mol. Microbiol. Biotechnol. 16, 109–123.

Geer, L.Y., Markey, S.P., Kowalak, J., Wagner, a, Xu, L., Maynard, M., Yang, D.M., Shi, X., Bryant, W., 2004. Open mass spectrometry search algorithm. J. Proteome Res. 3, 958–964.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A., 2011. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat. Biotechnol. 29, 644–652.

Gygi, S.P., Rochon, Y., Franza, B.R., Gygi, S.P., Rochon, Y., Franza, B.R., Aebersold, R., 1999. Correlation between protein and mRNA abundance in yeast correlation between protein and mRNA abundance in yeast. Mol. Cell. Biol. 19 (3), 1720–1730.

Hagel, J.M., Facchini, P.J., 2013. Benzylisoquinoline alkaloid metabolism: a century of discovery and a brave new world. Plant Cell. Physiol. 54, 647–672.

Hughes, C., Ma, B., Lajoie, G.A., 2010. De novo sequencing methods in proteomics. In: Hubbard, S.J., Jones, A.R. (Eds.), Proteome Bioinformatics, Methods in Molecular Biology. Springer, New York, pp. 105–121.

Illanes, A., Cauerhff, A., Wilson, L., Castro, G.R., 2012. Recent trends in biocatalysis engineering. Bioresour. Technol. 115, 48–57.

Jacobs, D.I., Gaspari, M., van der Greef, J., van der Heijden, R., Verpoorte, R., 2005. Proteome analysis of the medicinal plant *Catharanthus roseus*. Planta 221, 690–704.

Janssen, D.B., Dinkla, I.J.T., Poelarends, G.J., Terpstra, P., 2005. Bacterial degradation of xenobiotic compounds: evolution and distribution of novel enzyme activities. Environ. Microbiol. 7, 1868–1882.

Kim, S., Pevzner, P. a, 2014. MS–GF+ makes progress towards a universal database search tool for proteomics. Nat. Commun. 5, 5277.

Kirsch, R., Wielsch, N., Vogel, H., Svatoš, A., Heckel, D.G., Pauchet, Y., 2012. Combining proteomics and transcriptome sequencing to identify active plant-cell-wall-degrading enzymes in a leaf beetle. BMC Genomics 13, 587.

Kullman, S.W., Matsumura, F., 1996. Metabolic pathways utilized by phanerochaete chrysosporium for degradation of the cyclodiene pesticide endosulfan. Appl. Environ. Microbiol. 62, 593–600.
Li, S., Yang, X., Yang, S., Zhu, M., Wang, X., 2012. Technology prospecting on enzymes: application, marketing and engineering. Comput. Struct. Biotechnol. J. 2, e201209017.
McGettigan, P., 2013. Transcriptomics in the RNA-seq era. Curr. Opin. Chem. Biol. 17, 4–11.
Meyer, H.P., 2011. Sustainability and biotechnology. Org. Process Res. Dev. 15, 180–188.
Müller, R.-J., Schrader, H., Profe, J., Dresler, K., Deckwer, W.-D., 2005. Enzymatic degradation of poly(ethylene terephthalate): rapid hydrolyse using a hydrolase from T. fusca. Macromol. Rapid Commun. 26, 1400–1405.
Nordberg, H., Cantor, M., Dusheyko, S., Hua, S., Poliakov, A., Shabalov, I., Smirnova, T., Grigoriev, I.V., Dubchak, I., 2014. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic Acids Res. 42, D26–D31.
Ounaroon, A., Decker, G., Schmidt, J., Lottspeich, F., Kutchan, T.M., 2003. (R,S)-reticuline 7-*O*-methyltransferase and (R,S)-norcoclaurine 6-*O*-methyltransferase of *Papaver somniferum*—cDNA cloning and characterization of methyl transfer enzymes of alkaloid biosynthesis in opium poppy. Plant J. 36, 808–819.
Pappin, D.J.C., Creasy, D.M., Cottrell, J.S., 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data proteomics and 2-DE. Electrophoresis, 3551–3567.
Rabinovich, M.L., Bolobova, A.V., Vasil'chenko, L.G., 2004. Fungal decomposition of natural aromatic structures and xenobiotics: a review. Appl. Biochem. Microbiol. 40, 1–17.
Ribitsch, D., Acero, E.H., Greimel, K., Eiteljoerg, I., Trotscha, E., Freddi, G., Schwab, H., Guebitz, G.M., 2012. Characterization of a new cutinase from Thermobifida alba for PET-surface hydrolysis. Biocatal. Biotransform. 30, 2–9.
Robertson, D.E., Steer, B., 2004. Recent progress in biocatalyst discovery and optimization. Curr. Opin. Chem. Biol. 8, 141–149.
Saerens, K., Soetaert, W., 2011. Identification of key-genes from the sophorolipid biosynthetic pathway of Candida bombicola opens a new route to increased biosurfactant yields. Commun. Agricult. Appl. Biol. Sci., 65–68.
Saerens, K.M.J., Roelants, S.L.K.W., Van Bogaert, I.N.A., Soetaert, W., 2011. Identification of the UDP-glucosyltransferase gene UGTA1, responsible for the first glucosylation step in the sophorolipid biosynthetic pathway of Candida bombicola ATCC 22214. FEMS Yeast Res. 11, 123–132.
Saerens, K.M.J., Van Bogaert, I.N.A., Soetaert, W., 2015. Characterization of sophorolipid biosynthetic enzymes from Star merella bombicola. FEMS Yeast Res., 15.
Sarrouh, B., Santos, T.M., Miyoshi, A., Dias, R., Azevedo, V., 2012. Up-to-date insight on industrial enzymes applications and global market. J. Bioprocess. Biotechnol., 1–10, S4.
Schilmiller, A.L., Miner, D.P., Larson, M., McDowell, E., Gang, D.R., Wilkerson, C., Last, R.L., 2010. Studies of a biochemical factory: tomato trichome deep expressed sequence tag sequencing and proteomics. Plant Physiol. 153, 1212–1223.
Schittmayer, M., Birner-Gruenberger, R., 2012. Lipolytic proteomics. Mass Spectrom. Rev. 31, 570–582.
Schneider, T., Keiblinger, K.M., Schmid, E., Sterflinger-Gleixner, K., Ellersdorfer, G., Roschitzki, B., Richter, A., Eberl, L., Zechmeister-Boltenstern, S., Riedel, K., 2012. Who is who in litter decomposition? Metaproteomics reveals major microbial players and their biogeochemical functions. ISME J. 6, 1749–1762.
Spencer, J.F.T., Gorin, P.A.J., Tulloch, A.P., 1970. *Torulopsis bombicola* sp. n. Antonie Van Leeuwenhoek 36, 129–133.
Steinkellner, G., Gruber, C.C., Pavkov-Keller, T., Binter, A., Steiner, K., Winkler, C., Łyskowski, A., Schwamberger, O., Oberer, M., Schwab, H., Faber, K., Macheroux, P., Gruber, K., 2014. Identification of promiscuous ene-reductase activity by mining structural databases using active site constellations. Nat. Commun. 5, 1–9.
Tabb, D.L., Fernando, C.G., Chambers, M.C., 2008. MyriMatch: highly accurate tandem mass spectral peptide identificaiton by multivariate hypergeometric analysis. J. Proteome Res. 6, 654–661.
Tiwari, R., Singh, S., Singh, N., Adak, A., Rana, S., Sharma, A., Arora, A., Nain, L., 2014. Unwrapping the hydrolytic system of the phytopathogenic fungus *Phoma exigua* by secretome analysis. Process Biochem. 49, 1630–1636.
Uchiyama, T., Miyazaki, K., 2009. Functional metagenomics for enzyme discovery: challenges to efficient screening. Curr. Opin. Biotechnol. 20, 616–622.
Van Bogaert, I.N.A., Holvoet, K., Roelants, S.L.K.W., Li, B., Lin, Y.C., Van de Peer, Y., Soetaert, W., 2013. The biosynthetic gene cluster for sophorolipids: a biotechnological interesting biosurfactant produced by *Starmerella bombicola*. Mol. Microbiol. 88, 501–509.
Van Bogaert, I.N.A., Saerens, K., De Muynck, C., Develter, D., Soetaert, W., Vandamme, E.J., 2007. Microbial production and application of sophorolipids. Appl. Microbiol. Biotechnol. 76, 23–34.
Weiß, S., Lebuhn, M., Andrade, D., Zankel a. Cardinale, M., Birner-Gruenberger, R., Somitsch, W., Ueberbacher, B.J., Guebitz, G.M., 2013. Activated zeolite—suitable carriers for microorganisms in anaerobic digestion processes? Appl. Microbiol. Biotechnol. 97, 3225–3238.
Wilmes, P., Wexler, M., Bond, P.L., 2008. Metaproteomics provides functional insight into activated sludge wastewater treatment. PLoS One 3, e1778.
Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M.C., Zimmerman, L.J., Shaddox, K.F., Kim, S., Davies, S.R., Wang, S., Wang, P., Kinsinger, C.R., Rivers, R.C., Rodriguez, H., Townsend, R.R., Ellis, M.J.C., Carr, S., Tabb, a, Coffey, D.L., Slebos, R.J., Liebler, R.J.C., 2014. Proteogenomic characterization of human colon and rectal cancer. Nature 513, 382–387.
Zulak, K.G., Khan, M.F., Alcantara, J., Schriemer, D.C., Facchini, P.J., 2009. Plant defense responses in opium poppy cell cultures revealed by liquid chromatography–tandem mass spectrometry proteomics. Mol. Cell. Proteomics 8, 86–98.